University of Warsaw  
Faculty of Economic Sciences

Bartłomiej Kowalczuk

Album no. 372926

Michał ThorAlbum no. 361309

**Predicting bankruptcy among Polish companies using logistic regression and neural network**

Paper for Advanced Econometrics

on Data Science studies

Warsaw, may 2019

INTRODUCTION

Bankruptcies among companies were and still are widely discussed because of their impact on the economy. In fact bankruptcies have both good and bad influence on the macroeconomic situation – on one hand its’ occurrences result in loss of jobs and create market uncertainty (especially when bigger company defaults) but on the other hand it is an instrument of clearing the market from redundant firms. Although it may be useful, most of the market participants would rather know when certain company might “go down”. Focused efforts by economists and statisticians, that began in the thirties, effected in variety of tools designed to work as a precautionary mechanisms. However, prediction accuracy of these tools, especially econometric models, is often questioned. Considering Poland, most of the approaches from literature used to predict company bankruptcy are based on the discriminant analysis (such as Altman’s model) and logistic regression (such as Ohlson “O-score”). Also data used in those examples is not sampled – it is mainly based on finding every existing bankrupt company in the population and randomly attaching same number of healthy companies.

Main goal of our paper is to determine whether it is possible to create prediction model that gives reasonable company bankruptcy predictions on unbalanced data (in which bankrupt companies make only 3% of the whole dataset – as it is far more realistic scenario) in great advance (5 years prior to bankruptcy). Afterwards we try to compare it to model 1 year prior to default. If models appear to be applicable then we want to examine which accounting variables are the main determinants of these bankruptcies (and whether these determinants differ between models).

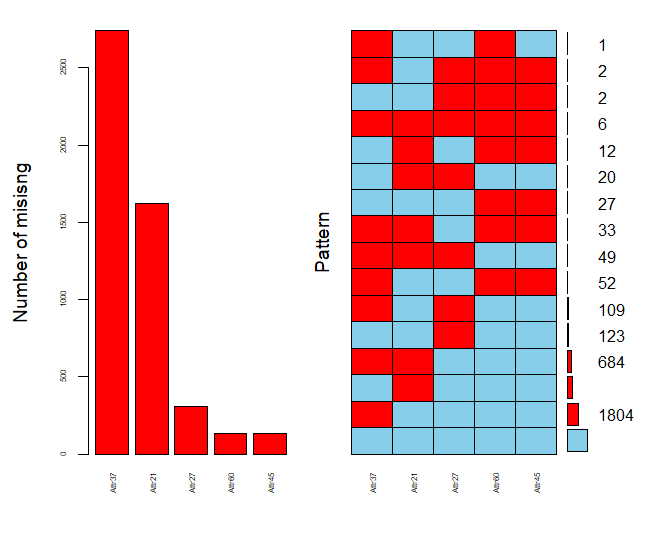
SECTION I

Preparing the dataset

Data which we used comes from a polish bankruptcies dataset on UCI Machine Learning Repository. Firstly we analyzed data for 1st year of forecasting period and therefore defaults mean that company default 5 years from that point. All of the basic 64 variables are transformed accounting data and various financial indicators (e.g. RoA, RoE, etc.). There were initially 7027 observations including 271 defaults which make for ~3.5% of the dataset.

Firstly we analyzed how many missing data are there in the set. Some variables were more prone to have NA’s but overall 3833 observations were not complete (including at least one missing variable). Number of missing for 5 most uncomplete variables are seen on the Figure 1.

Figure 1 - Variables with the most missing data

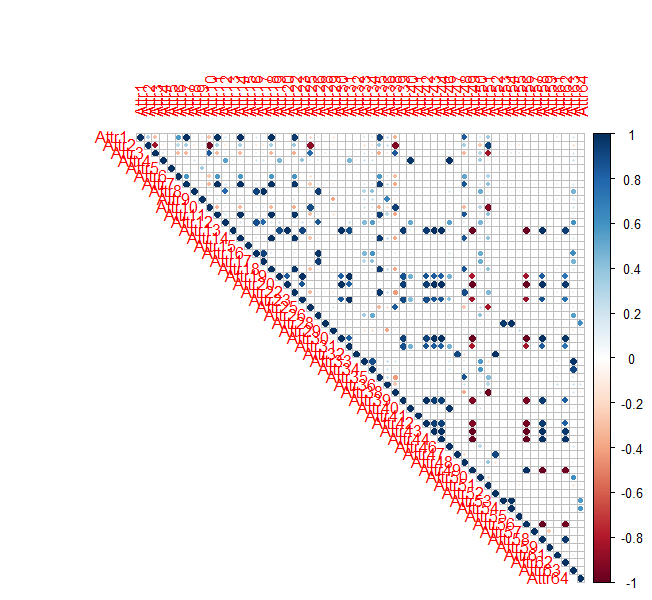


Wanting to retain most possible information while removing variables with unacceptable ratio of missing data, we introduced 5 factor variables indicating that there were missing data:

Subsequently we omitted columns with over 3% missing observations from further analysis. Since most of the missing came from the same observations we removed all of the (around 200 observations). Having in mind that retaining as much defaults as possible is crucial we find out that there are still around 3.5% of defaults. For some of the observations for which we already had the factor variable we decided to replace NA with median from the dataset (because they were single cases).

Next we analyzed correlation between variables. Correlation matrix can be seen on Figure 2.

Figure 2 - Correlation matrix for variables before removals

**

It can be seen that there are a lot of highly correlated variables. We filtered all the variables that are correlated with at least 2 other with correlation higher than 90% . Unfortunately it resulted in half of the variables being gone. Fortunately though it meant that there were almost no high correlated data and no missing observations at all. Correlation matrix for the cleaned data can be seen on Figure 3.

Figure 3 - Correlation matrix after removals

Obraz zawierający tekst, mapa

Opis wygenerowany automatycznie

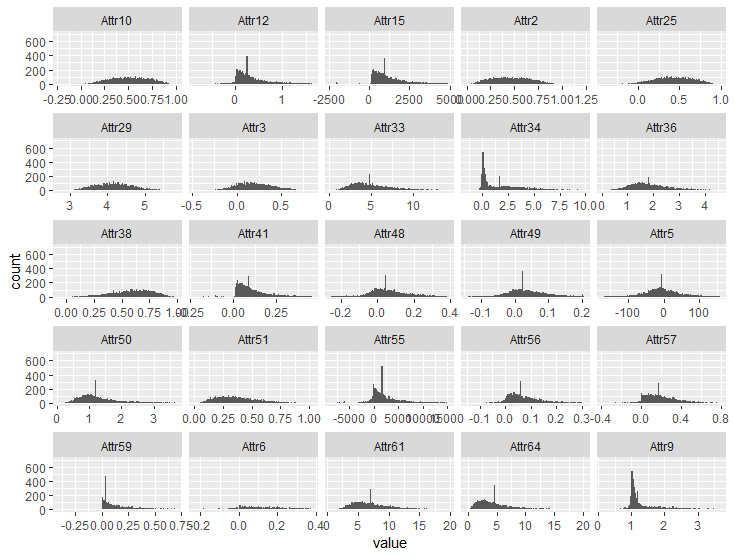
Correlations look much better, although some of them are still high. As we will be trying to retain as much variables as possible we will add those correlated variables as interactions when building model.

Further problem lies in the distribution of variables – there are A LOT of outliers in the data. With first estimations combined problems of high correlation and very high variance of most of the variables resulted in complete separation and Hauck Donner effect – instabilitiy of estimation in logisitc regression. To confront that problem we decided to replace some of the outliers with other values:

where: IQR – interquartile range

Therefore we are left with 4254 observations, while still retaining same percantage of defaults in dataset. On Figure 4 histograms of all (beside factor) variables (in Attr59 and Attr6 there are missing bins for zeros as they were larger than the scale – we shall emphasize on it below) are shown.

Figure 4 - Histograms for all variables



We can see that data is far from normalnny distributed and there are still values on the edges of distribution graph, but the improvement is undeniable. Also as mentioned below, Attr6 and Attr59 both have around 45% of zeros within its values but we decided to keep them in furhter analysis – maybe those with positive value of retained earnings/long-term liabilities are more prone to default.

Next step in our analysis is to transform some variables. After computing Information Values (IV) for all of the remaining variables, we decided that Attr12 should be binned as its suspiciously high predictive power (of 0.9) would work better when binned into two groups with borderline value of 0.06 which splits variable into groups with default rates of 14% and 2% respectively.

We also added interactions between correlated variables, interactions that would give us nominal values of accounting measures and some transformations to variables which had very high IQR.

SECTION II

Estimating models

With cleaned and prepared dataset with almost the same percentage of defaults as initial dataset we started estimating first, basic logarithmic regression model. At first we split our dataset into training and testing sets with a 70/30 proportion retaining default percentage in both subgroups. All of the models were validated using repeated k-fold cross validation for estimation of forecasts with 5 folds and 5 repeats.

First model was estimated on all variables with interactions. A lot of variables were statistically insignificant and so the AIC (Akaike’s Information Criterion) was 514.74. We set the cut-off point on 0.02, just to be sure that we are identifying as much bankrupts as we can. Its’ accuracy on test data was 79.76% but sensitivity (which in predicting bankruptcies should be our main goal) was only 83.72%. Estimated forecast sensitivity and specificity were From that point we firstly tried to remove those interactions and data transformation to see whether it will improve models’ accuracy.

Second model was estimated using same variables but we applied forward and back propagation (simultanously doing general-to-specific and specific-to-general methods) considering AIC as our point of reference (if removing certain variable improves AIC – remove it). We were left with 20 variables, although not all of them statistically significant at p-value = 0.05. AIC was 485.58, however accuracy of the model and its’ sensitivity did not improve – area under ROC curve and sensitivity became smaller.

At last we tried removing all those interactions and estimating model on only cleaned and prepared variables without interactions and transformed variables using AIC “both-side propagation” mentioned in second model. That way we obtained third model with AIC = 500.91 – worse than in second model, but its’ accuracy for cut-off point of 0.03 improved to 86.2% and, more importantly, sensitivity increased to 74.42%.

Comparison of the models’ coefficients estimations can be seen in the Appendix 1.

Comparison of the models’ accuracy is presented in Table 1 – AUC means area under curve. Underlined are the best values in each metric. Despite not being the most accurate overall we chose Model 3 as the best one because of the highest (by a margin) sensitivity.

Table 1 - Comparison of models' performance on test data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | AIC | Accuracy | Sensitivity | Specificity | auROC |
| Model 1 | 514.74 | 86.67 | 67.44 | 87.34 | 84.54 |
| Model 2 | 485.58 | 87.24 | 65.12 | 87.91 | 82.00 |
| **Model 3** | 500.91 | 86.20 | 74.42 | 86.61 | 85.78 |

Considering Model 3 as our best model, in Table 2 we present determinants along with their odds ratio – impact on the odds of bankruptcy. means that for one unit increase of a variable the odds of defaulting are lower. For we know that odds of defaulting increase by (or by ). For we know that that varaible does not have an impact on odds of going bankrupt.

Table 2 - Odds ratio for Model 3

|  |  |
| --- | --- |
| Variable | Odds ratio value |
| Attr2 | 190.987 |
| Attr3 | 0.145 |
| Attr6 | 0.031 |
| Attr9 | 1.472 |
| Attr10 | 55.012 |
| Attr12 | 0.251 |
| Attr15 | 1.000 |
| Attr33 | 1.100 |
| Attr38 | 0.001 |
| Attr49 | 530482.207 |
| Attr50 | 2.084 |
| Attr51 | 0.007 |
| Attr56 | 0.002 |
| Attr57 | 0.0416 |
| Attr651 | 0.427 |
| Attr661 | 311.164 |
| Attr681 | 4142383666.345 |

As we can see, variables have

Appendix 1 - Table of coefficients for Models 1,2,3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Results** | | | | | |
|  | Dependent variable:  class | | | | |
|  |  | |  | |  |
|  | Model 1 | Model 2 | | Model 3 | |
|  |  |  | |  | |
| Attr2 | 13.747 | 16.449\*\*\* | | 5.252\*\* | |
|  | (12.023) | (6.130) | | (2.196) | |
| Attr3 | 6.280 |  | | -1.931\* | |
|  | (13.339) |  | | (1.064) | |
| Attr5 | 0.002 |  | |  | |
|  | (0.005) |  | |  | |
| Attr6 | -3.443 | -4.405\*\* | | -3.458\* | |
|  | (2.248) | (1.891) | | (1.877) | |
| Attr9 | 0.614 |  | | 0.386\* | |
|  | (0.436) |  | | (0.211) | |
| Attr10 | 3.439 | 5.208\*\* | | 4.008\* | |
|  | (2.882) | (2.378) | | (2.358) | |
| Attr12 |  |  | | -1.384\*\* | |
|  |  |  | | (0.668) | |
| Attr15 | 0.0003\* | 0.0003\*\* | | 0.0002\* | |
|  | (0.0001) | (0.0001) | | (0.0001) | |
| Attr25 | -2.468 |  | |  | |
|  | (3.628) |  | |  | |
| Attr29 | 0.542 |  | |  | |
|  | (1.032) |  | |  | |
| Attr33 | 0.233\*\* | 0.174\*\* | | 0.095 | |
|  | (0.102) | (0.068) | | (0.065) | |
| Attr34 | 0.124 |  | |  | |
|  | (0.110) |  | |  | |
| Attr36 | -0.582 |  | |  | |
|  | (0.363) |  | |  | |
| Attr38 | -6.445\*\* | -6.460\*\*\* | | -7.035\*\*\* | |
|  | (2.787) | (2.096) | | (2.059) | |
| Attr41 | 0.399 |  | |  | |
|  | (1.869) |  | |  | |
| Attr48 | 1.501 |  | |  | |
|  | (2.534) |  | |  | |
| Attr49 | 0.523 |  | | 13.182\*\*\* | |
|  | (5.439) |  | | (3.251) | |
| Attr50 | 0.588 | 0.805\* | | 0.734\* | |
|  | (0.498) | (0.444) | | (0.396) | |
| Attr51 | 10.517 |  | | -4.909\*\*\* | |
|  | (8.720) |  | | (1.823) | |
| Attr55 | -0.001\* | -0.001 | |  | |
|  | (0.001) | (0.001) | |  | |
| Attr56 | -9.862\*\*\* | -7.761\*\*\* | | -6.295\*\* | |
|  | (2.996) | (2.590) | | (2.464) | |
| Attr57 | -2.839\*\* | -2.925\*\*\* | | -3.180\*\*\* | |
|  | (1.173) | (0.938) | | (1.066) | |
| Attr59 | 0.504 |  | |  | |
|  | (0.980) |  | |  | |
| Attr61 | 0.017 |  | |  | |
|  | (0.160) |  | |  | |
| Attr64 | -0.175 | -0.067 | |  | |
|  | (0.153) | (0.050) | |  | |
| Attr651 | -1.536\*\*\* | -1.292\*\*\* | | -0.851\*\* | |
|  | (0.452) | (0.385) | | (0.349) | |
| Attr661 | 13.491\*\*\* | 6.052\*\*\* | | 5.740\*\*\* | |
|  | (4.376) | (0.429) | | (0.393) | |
| Attr671 | 0.647 |  | |  | |
|  | (0.990) |  | |  | |
| Attr681 | 22.791 | 21.413 | | 22.145 | |
|  | (678.988) | (728.862) | | (670.732) | |
| Attr12\_BINNED | 0.350 |  | |  | |
|  | (0.405) |  | |  | |
| I(Attr22) | 0.512 | -6.331 | |  | |
|  | (8.316) | (4.367) | |  | |
| I(Attr253) | 9.247\*\* | 9.285\*\* | |  | |
|  | (4.513) | (3.656) | |  | |
| I(log1p(Attr61)) | -0.437 |  | |  | |
|  | (1.421) |  | |  | |
| I(log1p(Attr64)) | 0.848 |  | |  | |
|  | (1.308) |  | |  | |
| Attr2:Attr29 | -0.326 |  | |  | |
|  | (1.372) |  | |  | |
| Attr661:Attr12\_BINNED | 0.579 |  | |  | |
|  | (0.875) |  | |  | |
| Attr38:Attr661 | -6.905 |  | |  | |
|  | (4.407) |  | |  | |
| Attr9:I(exp(Attr29)) | -0.001 |  | |  | |
|  | (0.004) |  | |  | |
| Attr2:Attr51 | -17.778\* | -5.018\*\* | |  | |
|  | (10.777) | (2.240) | |  | |
| Attr2:Attr661 | -6.253 |  | |  | |
|  | (4.023) |  | |  | |
| Attr2:Attr3 | -4.578 |  | |  | |
|  | (13.802) |  | |  | |
| Attr3:Attr10 | 11.018 | 14.307\*\* | |  | |
|  | (13.591) | (7.275) | |  | |
| Attr3:Attr25 | -4.844 |  | |  | |
|  | (5.074) |  | |  | |
| Attr3:Attr38 | -13.151 | -13.189\*\* | |  | |
|  | (9.531) | (6.593) | |  | |
| Attr25:Attr38 | 0.253 | -6.054\*\* | |  | |
|  | (7.983) | (2.660) | |  | |
| Attr49:Attr56 | 93.948\*\*\* | 93.214\*\*\* | |  | |
|  | (26.153) | (19.227) | |  | |
| Attr29:Attr55 | 0.0002\* | 0.0002 | |  | |
|  | (0.0001) | (0.0001) | |  | |
| Constant | -12.625\* | -9.171\*\*\* | | -3.952\* | |
|  | (6.861) | (3.282) | | (2.311) | |
|  |  |  | |  | |
| Observations | 2,979 | 2,979 | | 2,979 | |
| Log Likelihood | -210.369 | -219.791 | | -232.456 | |
| Akaike Inf. Crit. | 514.738 | 485.583 | | 500.912 | |
|  |  |  | |  | |
|  | | | | | |
| Note: | \*p<0.1; | \*\*p<0.05 | | \*\*\*p<0.01 | |